Mixing Time Estimation in Ergodic Markov Chains from a Single Trajectory with Contraction Methods

Geoffrey Wolfer GEOFFREY@POST.BGU.AC.IL
Department of Computer Science
Ben-Gurion University of the Negev
Israel

Editors: Aryeh Kontorovich and Gergely Neu

Abstract

The mixing time t_{mix} of an ergodic Markov chain measures the rate of convergence towards its stationary distribution $\boldsymbol{\pi}$. We consider the problem of estimating t_{mix} from one single trajectory of m observations $(X_1, \dots, X_m)$, in the case where the transition kernel $\boldsymbol{M}$ is unknown, a research program started by Hsu et al. (2015). The community has so far focused primarily on leveraging spectral methods to estimate the *relaxation time* t_{rel} of a *reversible* Markov chain as a proxy for t_{mix}. Although these techniques have recently been extended to tackle non-reversible chains, this general setting remains much less understood. Our new approach based on contraction methods is the first that aims at directly estimating t_{mix} up to multiplicative small universal constants instead of t_{rel}. It does so by introducing a generalized version of Dobrushin's contraction coefficient κ_{gen}, which is shown to control the mixing time regardless of reversibility. We subsequently design fully data-dependent high confidence intervals around κ_{gen} that generally yield better convergence guarantees and are more practical than state-of-the-art.

Keywords: Ergodic Markov chain, mixing time, Dobrushin contraction coefficient

1. Introduction

The topic of this work is the construction of a non-trivial high confidence interval around the mixing time of a finite state ergodic Markov chain, when one is only allowed to observe a single long trajectory of states $X_1, X_2, \dots, X_m$, i.e. does not have access to a restart mechanism. The problem is motivated by PAC-type learning problems that assume data sampled from a Markovian process, where generalization guarantees oftentimes involve the *a priori* unknown mixing properties of the chain. Other applications are in MCMC diagnostics for non-reversible Markov chains, that may enjoy better mixing properties or asymptotic variance than their reversible counterparts, or in the field of reinforcement learning, where bounds on the mixing time are routinely assumed. We invite the reader to the *related work* sections of Hsu et al. (2019); Wolfer and Kontorovich (2019) for a complete set of references to the aforementioned problems and additional motivation.

Main contributions.

- In Section 2, in lieu of the (pseudo-)relaxation time t_{rel}, we propose a new proxy for the mixing time based on a *contraction coefficient* κ_{gen} that generalizes Dobrushin's, and in particular, does not require reversibility. We show in Theorem 1 that this quantity controls the mixing time up to multiplicative universal constants – which are small and given at (10) –

such that contrary to the relaxation time, it is not subject to the gap mentioned at (5). Namely,

$$\frac{1}{1-\kappa_{\mathsf{gen}}} = \Theta(t_{\mathsf{mix}}).$$

- In Section 3.1, we design fully empirical confidence intervals around κ_{gen} that in the general (non-reversible) setting are thinner, and considerably more practical than their spectral state-of-the-art counterparts: For a chain on d states and a chosen parameter $S \in \mathbb{N}$, our estimator $\widehat{\kappa}_{\mathsf{gen}[S]}$ is such that

$$\left|\widehat{\kappa}_{\mathsf{gen}[S]} - \kappa_{\mathsf{gen}}\right| \leq \tilde{\mathcal{O}}\left(\frac{1}{S} + \max_{\ell \in [S]} \left\{\frac{1}{\ell}\sqrt{\frac{d}{N_{\min}^{(\ell)}}}\right\}\right),$$

  where $N_{\min}^{(\ell)}$ is the *least number of visits for the ℓ-skipped chain*, a fully observable quantity defined at (11). Additionally, the analysis leading to the confidence intervals is of an arguably much simpler nature than that of Wolfer and Kontorovich (2019).

- In Section 3.2, for a d state Markov chain with minimum stationary probability $\pi_\star$ (definition at (2)), we further deduce point estimators for estimating κ_{gen} down to absolute error ε (Theorem 5), with sample complexity $m_+ = \tilde{\mathcal{O}}\left(\frac{1}{\pi_\star}\max\left\{t_{\mathsf{mix}}, \frac{d}{\varepsilon^2}\right\}\right)$ and relative error ε (Theorem 6), for a trajectory length of $m_\times = \tilde{\mathcal{O}}\left(\frac{dt_{\mathsf{mix}}^2}{\pi_\star \varepsilon^2}\right)$, offering better guarantees than the one of Wolfer and Kontorovich (2019) for the non-trivial classes of slow mixing chains ($t_{\mathsf{mix}} > d$), and chains whose stationary distribution is not close to being uniform (see Remark 8).

Notation and setting. The set $\mathbb{N}$ will refer to $\{1, 2, 3, \ldots\}$ and for $n \in \mathbb{N}$, we write $[n] = \{1, 2, 3, \ldots, n\}$. Let Ω a set such that $|\Omega| = d < \infty$, and define Δ_Ω the simplex of all distributions – seen as row vectors – over Ω. For $(\boldsymbol{\mu}, \boldsymbol{\nu}) \in \Delta_\Omega^2$, we define the *total variation distance* in terms of the ℓ_1 norm:

$$\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_{\mathsf{TV}} \doteq \frac{1}{2}\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_1 . \tag{1}$$

We consider *time-homogeneous Markov chains*

$$X_1, X_2, \ldots, X_t, \ldots \sim (\boldsymbol{\mu}, \boldsymbol{M})$$

with *initial distribution* $\boldsymbol{\mu} \in \Delta_\Omega$, and *row-stochastic transition matrix* $\boldsymbol{M} \colon \Omega \times \Omega \to [0, 1]$. We say that a Markov chain $(\boldsymbol{\mu}, \boldsymbol{M})$ is *ergodic* when $\boldsymbol{M}$ is a *primitive matrix*, i.e. $\exists p \in \mathbb{N}, \boldsymbol{M}^p > 0$ entry-wise. In this case, $\boldsymbol{M}$ has a unique *stationary distribution* $\boldsymbol{\pi}$ such that $\boldsymbol{\pi}\boldsymbol{M} = \boldsymbol{\pi}$, the chain is known to converge to $\boldsymbol{\pi}$, and the *minimum stationary probability*

$$\pi_\star \doteq \min_{i \in \Omega} \boldsymbol{\pi}(i) \tag{2}$$

is such that $\pi_\star > 0$. We measure distance to stationarity in total variation,

$$h(t) \doteq \sup_{\boldsymbol{\mu} \in \Delta_\Omega} \left\|\boldsymbol{\mu}\boldsymbol{M}^t - \boldsymbol{\pi}\right\|_{\mathsf{TV}} . \tag{3}$$

For $\xi \in (0, 1/2)$, the *mixing time* of $\boldsymbol{M}$ is defined by

$$t_{\mathsf{mix}}(\xi) \doteq \arg\min_{t \in \mathbb{N}} \{h(t) < \xi\}, \tag{4}$$

and by convention $t_{\mathsf{mix}} \doteq t_{\mathsf{mix}}(1/4)$. The reader is referred to Levin et al. (2009, Chapter 4) for a more detailed introduction to Markov chain mixing. We will use the standard $\mathcal{O}$ and Θ notations, and $\tilde{\mathcal{O}}, \tilde{\Theta}$ when logarithmic dependencies in any natural parameter are omitted, and for $x \in \mathbb{R}_+$, we will use $\widetilde{\ln}\ x$ as a shorthand for $\ln x \ln\ln x$. The definition of elements specific to contraction methods is deferred to Section 2 for a clearer exposition.

Related work. Research has so far mostly focused on leveraging spectral methods for estimating the *relaxation time* t_{rel} of a *reversible* (Levin et al., 2009, Section 1.6) chain as an approximation of t_{mix} (Hsu et al., 2015; Levin and Peres, 2016; Hsu et al., 2019; Combes and Touati, 2019; Qin et al., 2019). Indeed, in this setting, t_{rel}, defined as the inverse of the *absolute spectral gap* $\gamma_\star$, is known to be related – see (5) – to the mixing time up to a logarithmic correction (Levin et al., 2009, Theorem 12.4). The problem, therefore, reduces to estimating the second largest eigenvalue in magnitude. Moreover, as reversibility and self-adjointness of the Markov operator are equivalent notions, dimension-free perturbation eigenvalue bounds – namely Weyl's inequality – are available to efficiently estimate its spectrum, which is the subject of Hsu et al. (2015); Levin and Peres (2016); Hsu et al. (2019). In their work Combes and Touati (2019), offer a different perspective on the problem by putting the emphasis on computational complexity, invoking *power methods* and *upper confidence interval* techniques to design a more space-efficient estimator. Finally, Qin et al. (2019) explore the case of general state spaces, for kernels that are *trace-class* operators (compact with summable eigenvalues). Although broad collections of chains are known to be reversible such as random walks on graphs or birth and death processes, this assumption is a strong restriction on the class of chains that can be treated, and our approach compares favorably with this body of work in as much as it removes this requirement.

One exception to the above list is Wolfer and Kontorovich (2019) that extends the estimation results to the *non-reversible* setting, by estimating the pseudo-relaxation time (Kamath and Verdú, 2016, (16)). More specifically, even in the absence of reversibility, it was shown in Paulin (2015, Proposition 3.4) that a related quantity, the inverse of the *pseudo-spectral gap*

$$\gamma_{\mathsf{ps}} \doteq \max_{k \in \mathbb{N}} \left\{ \gamma((\boldsymbol{M}^\dagger)^k \boldsymbol{M}^k)/k \right\},$$

where $\boldsymbol{M}^\dagger$ is the time-reversal of $\boldsymbol{M}$, still traps the mixing time up to a logarithmic correction. Wolfer and Kontorovich (2019) carry out the analysis of estimating this quantity, with a scheme that consists in observing multi-step chains forward and backward in time. They show that it is enough to explore a finite set of skipping rates, and recover a consistent estimator that still enjoys spectral stability, and converges to arbitrary precision with a trajectory length polynomial in the natural parameters $d, \pi_\star, \varepsilon, t_{\mathsf{mix}}$.

An inherent drawback of *all* previous approaches, however, is the existence of a known gap between the (pseudo-)relaxation time and t_{mix} that depends on $\pi_\star$ or d (Levin et al., 2009, Theorem 12.4), (Paulin, 2015, Proposition 3.4), (Jerison, 2013, Theorem 1.2). We can summarize these results as

$$c_1 \cdot (t_{\mathsf{rel}} - 1) \leq t_{\mathsf{mix}} \leq c_2 \cdot \min\left\{ d, \ln \frac{1}{\pi_\star} \right\} t_{\mathsf{rel}}, \text{ where } (c_1, c_2) \in \mathbb{R}_+^2. \tag{5}$$

Moreover, it is known that this gap cannot generally be closed (Jerison, 2013), so that estimation of t_{rel} down to arbitrary error still will not yield an accurate estimate for t_{mix} as $d \to \infty$. This limitation is the motivation behind our search for a new, tighter proxy. Although the task of estimating t_{mix} directly is currently believed to be more challenging than t_{rel} as raised in the concluding remarks of Combes and Touati (2019, Conclusion), no rigorous or quantitative comparison is known in terms of statistical complexity. This work therefore also initiates the investigations towards answering this question. Comparisons of convergence rates with the state-of-the-art point empirical confidence intervals and estimators are respectively carried out at Remark 4 and Remark 8.

Finally, as raised in Combes and Touati (2019), there exists an interesting trade-off between computational complexity and statistical accuracy. This work is more concerned with the latter, such that the designed procedures will be applicable for medium-sized state spaces, with computational complexities of the same order to that of Wolfer and Kontorovich (2019).

2. Generalized contraction coefficient

For a Markov chain $\boldsymbol{M}$, the *Dobrushin contraction* coefficient, also known as *Dobrushin ergodic* coefficient (Dobrushin, 1956), (Brémaud, 1999, Definition 7.1) is defined by

$$\kappa \doteq \max_{(i,j)\in\Omega^2} \|\boldsymbol{M}(i,\cdot) - \boldsymbol{M}(j,\cdot)\|_{\mathsf{TV}}\,, \tag{6}$$

where the term *contraction* refers to the property (Brémaud, 1999, Corollary 7.1) that $\forall(\boldsymbol{\mu},\boldsymbol{\nu}) \in \Delta_\Omega^2$,

$$\|(\boldsymbol{\mu}-\boldsymbol{\nu})\boldsymbol{M}\|_{\mathsf{TV}} \le \kappa\,\|\boldsymbol{\mu}-\boldsymbol{\nu}\|_{\mathsf{TV}}\,. \tag{7}$$

Contraction in the sense of Dobrushin is a special case of *coarse Ricci curvature* (Ollivier, 2009), where the metric taken on Ω is the *discrete metric*, and the *Wasserstein distance* between distributions reduces to total variation. In the case where $\kappa < 1$, the Bubley-Dyer path coupling bound (Bubley and Dyer, 1997) gives an upper bound on mixing time

$$t_{\mathsf{mix}}(\xi) \le \frac{\ln \xi}{\ln(1-\kappa)}$$

Unfortunately, there exists a large subset of ergodic chains such that $\kappa = 1$, and for which this direct method fails to yield convergence rates. To overcome this limitation, we consider multi-step chains, where for $s \in [m-1]$,

$$X_1, X_{1+s}, X_{1+2s}, \ldots, X_{1+\lfloor (m-1)/s\rfloor s} \sim (\boldsymbol{\mu}, \boldsymbol{M}^s),$$

and define the contraction coefficient of the chain with skipping rate s to be

$$\kappa_s \doteq \max_{(i,j)\in\Omega^2} \|\boldsymbol{M}^s(i,\cdot) - \boldsymbol{M}^s(j,\cdot)\|_{\mathsf{TV}}\,. \tag{8}$$

We then introduce a *generalized contraction coefficient* κ_{gen} of the ergodic chain $\boldsymbol{M}$:

$$\kappa_{\mathsf{gen}} \doteq 1 - \max_{s\in\mathbb{N}} \left\{ \frac{1-\kappa_s}{s} \right\}, \tag{9}$$

and write s_{gen} the smallest integer [1] such that $\kappa_{\mathsf{gen}} = 1 - \frac{1-\kappa_{s_{\mathsf{gen}}}}{s_{\mathsf{gen}}}$. This quantity is derived in a similar spirit as Paulin (2015) defined the pseudo-spectral gap of an ergodic Markov chain. Even in the case where $\kappa = 1$, we now formalize in Theorem 1 the fact that $\kappa_{\mathsf{gen}} < 1$ always holds, as $\frac{1}{1-\kappa_{\mathsf{gen}}}$ traps t_{mix} up to universal constants.

Theorem 1 *Let $\xi \in (0, 1/2)$, and $\boldsymbol{M}$ ergodic with mixing time $t_{\mathsf{mix}}(\xi)$, then*

$$\frac{1-2\xi}{1-\kappa_{\mathsf{gen}}} \leq t_{\mathsf{mix}}(\xi) \leq \frac{1+\ln 1/\xi}{1-\kappa_{\mathsf{gen}}},$$

where κ_{gen} is defined at (9), and in particular,

$$\frac{1/2}{1-\kappa_{\mathsf{gen}}} \leq t_{\mathsf{mix}} \leq \frac{1+\ln 4}{1-\kappa_{\mathsf{gen}}}. \tag{10}$$

We point out that although we could not find any reference to the quantity at (9), or to Theorem 1, considering multi-step contractions to study concentration or mixing properties of chains is not a novel idea in itself; see for instance Dyer et al. (2001); Luczak (2008); Paulin (2016).

3. Statistical estimation of the mixing time from a single trajectory

This section is devoted to the analysis of the statistical complexity of estimating the mixing time of an ergodic chain from one single long draw of observations (no restart mechanism). Section 2 introduced κ_{gen} [defined at (9)] as a tighter proxy for t_{mix} [as shown by Theorem 1] and allows for a reduction of the estimation problem. In Section 3.1 we construct fully empirical high-confidence intervals around κ_{gen}. We further derive point estimators in Section 3.2 and analyze their finite sample convergence properties both in *absolute* (Theorem 5) and *relative* (Theorem 6) error.

3.1. Fully empirical confidence intervals

For a confidence parameter δ, and a trajectory $X_1, \ldots, X_m$, our goal is to construct a non-trivial interval $I_{\kappa_{\mathsf{gen}}} = (\kappa_{\mathsf{gen,lb}}, \kappa_{\mathsf{gen,ub}})$ such that

$$\mathbb{P}\left(\kappa_{\mathsf{gen}} \in I_{\kappa_{\mathsf{gen}}}\right) \geq 1 - \delta.$$

Our estimator will be a truncated plug-in version of κ_{gen}, where we only explore a prefix $[S]$ of the integers, a similar idea as employed in Wolfer and Kontorovich (2019) for estimating the pseudo-spectral gap (Paulin, 2015, Section 3.1). For a chain with skipping rate s, we define the following random variables,

$$\begin{aligned} N_i^{(s)} &\doteq \sum_{t=1}^{\lfloor (m-1)/s \rfloor} \mathbf{1}\left\{X_{1+s(t-1)} = i\right\}, \qquad N_{\min}^{(s)} \doteq \min_{i \in \Omega} N_i^{(s)}, \\ N_{ij}^{(s)} &\doteq \sum_{t=1}^{\lfloor (m-1)/s \rfloor} \mathbf{1}\left\{X_{1+s(t-1)} = i, X_{1+st} = j\right\}, \end{aligned} \tag{11}$$

1. The existence of s_{gen} is guaranteed by the observation that $s \mapsto \frac{1-\kappa_s}{s} \in (0, \frac{1}{s})$.

and construct an estimator for the multi-step kernel $\boldsymbol{M}^s$ and its contraction coefficient,

$$\widehat{\boldsymbol{M}}^{(s)} \doteq \sum_{(i,j)\in\Omega^2} \frac{N_{ij}^{(s)}}{N_i^{(s)}} \mathbf{1}\left\{N_i^{(s)} > 0\right\} \boldsymbol{e}_i \otimes \boldsymbol{e}_j, \qquad \widehat{\kappa}_s \doteq \kappa\left(\widehat{\boldsymbol{M}}^{(s)}\right), \tag{12}$$

where $\boldsymbol{e}_i$ is the ith coordinate basis vector and $\otimes$ denotes the standard tensor product. When $s = 1$, we will omit subscript or superscript and write respectively $N_i, N_{ij}, N_{\min}, \widehat{\boldsymbol{M}}$ and $\widehat{\kappa}$. Finally, the estimator for κ_{gen} parametrized by an integer S is

$$\widehat{\kappa}_{\mathsf{gen}[S]} \colon \Omega^m \to (0,1),\ \boldsymbol{X} \mapsto 1 - \max_{s\in[S]} \left\{\frac{1-\widehat{\kappa}_s(\boldsymbol{X})}{s}\right\}.$$

Theorem 2 *Let $\delta \in (0,1)$, $S \in \mathbb{N}$, and $X_1, \ldots, X_m \sim \boldsymbol{M}$, then with probability $1-\delta$,*

$$\left|\widehat{\kappa}_{\mathsf{gen}[S]} - \kappa_{\mathsf{gen}}\right| \le \frac{1}{S} + \sqrt{d} \max_{s\in[S]} \left\{\frac{\mathcal{L}_s}{s\sqrt{N_{\min}^{(s)}}}\right\},$$

where $\mathcal{L}_s = \mathcal{O}\left(\ln\left(\frac{dS\ln m/s}{\delta}\right)\right)$, and $N_{\min}^{(s)}$ is defined at (11).

Remark 3 *As we choose to carry out our analysis with unsmoothed estimators, we see that the confidence bounds can be ill-defined for short trajectories. A slight modification of the proofs with a smoothing parameter, for example, analyzing*

$$\widehat{\boldsymbol{M}}^{(\lambda,s)} \doteq \sum_{(i,j)\in\Omega^2} \frac{N_{ij}^{(s)} + \lambda}{N_i^{(s)} + d\lambda} \boldsymbol{e}_i \otimes \boldsymbol{e}_j,$$

with $\lambda > 0$ instead, can yield intervals that are well defined almost surely. This would, however, clutter the analysis while offering only incremental improvement.

Remark 4 *Not only is the interval at Theorem 2 far more* user-friendly *than the one designed around the pseudo-spectral gap in Wolfer and Kontorovich (2019, Theorem 8), it is also much narrower. Denoting by $\approx$ a rough estimate of the rate at which we expect the intervals to decay in width,*

$$\left|\widehat{\kappa}_{\mathsf{gen}[S]} - \kappa_{\mathsf{gen}}\right| \approx \frac{1}{S} + \sqrt{\frac{d}{\pi_\star m}}, \tag{13}$$

whereas the known intervals around the pseudo-spectral gap γ_{ps} of the estimator $\widehat{\gamma}_{\mathsf{ps}[S]}$ defined by Wolfer and Kontorovich could generally decay as slowly as

$$\left|\widehat{\gamma}_{\mathsf{ps}[S]} - \gamma_{\mathsf{ps}}\right| \approx \frac{1}{S} + \frac{1}{\pi_\star^{3/2}}\sqrt{\frac{d}{m}}\left(\sqrt{d} + \frac{1}{\gamma_{\mathsf{ps}}}\right).$$

We end this section with a short discussion on the choice of S, that is missing from Wolfer and Kontorovich (2019). Assuming a $1-\delta/2$ confidence interval $I_{\pi_\star} = (\pi_{\star,\mathsf{lb}}, \pi_{\star,\mathsf{ub}})$ around $\pi_\star$, for instance employing the estimation procedure of Hsu et al. (2015), then a practical choice of S for balancing the two terms at (13) is

$$S \approx n \vee \sqrt{m(\pi_{\star,\mathsf{lb}} \wedge 1/d)/d}, \tag{14}$$

where n is a small arbitrary integer. In other words, it is reasonable to wait for the trajectory length to be of the order of the square of the state space size before starting to explore larger skipping rates.

3.2. Point estimator for κ_{gen}

For chosen precision ε and confidence δ parameters, we construct a point estimator down to absolute error, where the algorithm only needs knowledge of d and ε, δ in order to run.

Theorem 5 (Point estimator for κ_{gen} (absolute error)) *Let $(\varepsilon, \delta) \in (0,1)^2$, and let*

$$X_1, X_2, \ldots, X_m \sim (\boldsymbol{\mu}, \boldsymbol{M})$$

an unknown ergodic Markov chain with minimum stationary probability $\pi_\star$, and generalized contraction coefficient κ_{gen}. There exists an estimation procedure $\widehat{\kappa}^{+}_{\mathsf{gen}} \colon \Omega^m \to (0,1)$ such that for

$$m \geq c \frac{\mathcal{L}}{\pi_\star} \max\left\{ \frac{1}{1-\kappa_{\mathsf{gen}}}, \frac{d}{\varepsilon^2} \right\},$$

$\left|\widehat{\kappa}^{+}_{\mathsf{gen}} - \kappa_{\mathsf{gen}}\right| < \varepsilon$ *holds with probability at least $1-\delta$, where $\mathcal{L} = \mathcal{O}\left(\widetilde{\ln}\left(\frac{d}{\delta\varepsilon\pi_\star}\right)\right)$, and c is a universal constant.*

From the proof of Theorem 5, we observe that, perhaps surprisingly, for a contracting chain, i.e. $\kappa = 1-\alpha$ with $\alpha > 0$, the statistical difficulty of estimating α is of the same order (ignoring logarithmic factors) as that of estimating κ_{gen}, while only providing with – generally sub-optimal– upper bounds on the mixing time. One remaining question is the necessity of the dependency in d. A heuristic argument based on the results of Jiao et al. (2018) would seem to imply that any technique basing itself solely on the definition of a contraction coefficient – boiling down to estimating ℓ_1 distances of $[d]$ supported distributions – would necessarily have a statistical dependency in the support size.

We now show that it is also possible to construct an algorithm that outputs an estimate of $1-\kappa_{\mathsf{gen}}$ with relative error ε. For this goal, the algorithm needs to explore at least the first $S = \Theta\left(\frac{1}{\varepsilon(1-\kappa_{\mathsf{gen}})}\right)$, involving the unknown quantity κ_{gen}. The solution is an adaptive argument that is fleshed out in Section 4.4.

Theorem 6 (Point estimator for $1-\kappa_{\mathsf{gen}}$ (relative error)) *Let $(\varepsilon, \delta) \in (0,1)^2$, and let*

$$X_1, X_2, \ldots, X_m \sim (\boldsymbol{\mu}, \boldsymbol{M})$$

an unknown ergodic Markov chain with minimum stationary probability $\pi_\star$, and generalized contraction coefficient κ_{gen}. There exists an estimation procedure $\widehat{\kappa}^{\times}_{\mathsf{gen}} \colon \Omega^m \to (0,1)$ such that for

$$m \geq c \frac{\mathcal{L} d}{\pi_\star (1-\kappa_{\mathsf{gen}})^2 \varepsilon^2},$$

$\left|\frac{1-\widehat{\kappa}^{\times}_{\mathsf{gen}}}{1-\kappa_{\mathsf{gen}}} - 1\right| < \varepsilon$ *holds with probability at least $1-\delta$, where $\mathcal{L} = \mathcal{O}\left(\widetilde{\ln}\left(\frac{d}{\delta\varepsilon\pi_\star(1-\kappa_{\mathsf{gen}})}\right)\right)$, and c is a universal constant.*

Corollary 7 (to Theorem 6) *Let $\delta \in (0,1)$, and let $X_1, X_2, \ldots, X_m \sim (\boldsymbol{\mu}, \boldsymbol{M})$ an unknown ergodic Markov chain with minimum stationary probability $\pi_\star$, and mixing time t_{mix}. There exists an estimation procedure $\hat{t}_{\mathsf{mix}} \colon \Omega^m \to \mathbb{N}$ such that for $m \geq c\frac{\mathcal{L} d t_{\mathsf{mix}}^2}{\pi_\star}$, $\frac{1}{3} t_{\mathsf{mix}} \leq \hat{t}_{\mathsf{mix}} \leq 3 t_{\mathsf{mix}}$, holds with probability at least $1-\delta$, where $\mathcal{L} = \mathcal{O}\left(\widetilde{\ln}\left(\frac{t_{\mathsf{mix}} d}{\delta\pi_\star}\right)\right)$, and c is a universal constant.*

Remark 8 *Although in principle, direct comparison of point estimators with prior research is not possible as all previous work focused on t_{rel}, we treat for now the question as if estimation of κ_{gen} or t_{mix} directly is not harder, and focus on relative error. In the general (non-reversible) setting, the only known finite sample upper bound (Wolfer and Kontorovich, 2019, Theorem 3) for estimating the pseudo-relaxation time is of*

$$m_{\times} = \tilde{\mathcal{O}}\left(\frac{t_{\mathsf{rel}}^2}{\pi_\star \varepsilon^2} \max\left\{t_{\mathsf{rel}}, \beta(\boldsymbol{\pi}) \min\left\{\beta(\boldsymbol{\pi}), d\right\}\right\}\right),$$

where $\beta(\boldsymbol{\pi}) \doteq \max_{(i,j)\in\Omega^2}\left\{\frac{\boldsymbol{\pi}(i)}{\boldsymbol{\pi}(j)}\right\}$ measures how far $\boldsymbol{\pi}$ is from being uniform, and $1 \le \beta(\boldsymbol{\pi}) \le 1/\pi_\star$. From Theorem 6, it is possible to estimate $1-\kappa_{\mathsf{gen}}$ with $m_{\times} = \tilde{\mathcal{O}}\left(\frac{dt_{\mathsf{mix}}^2}{\pi_\star \varepsilon^2}\right)$, so that for the two classes of slow mixing chains ($t_{\mathsf{mix}} > d$) and chains with a stationary distribution $\boldsymbol{\pi}$ such that $\beta(\boldsymbol{\pi}) > \sqrt{d}$, our result dominates, showing that the two methods offer complementary convergence rates.

4. Proofs

4.1. Proof of Theorem 1.

The proof is standard. See for example Paulin (2015, Section 5.2) for a similar technique. We will first bound the distance to stationarity $h(t)$ for a given $t > s_{\mathsf{gen}}$,

$$\begin{aligned}
h(t) &\doteq \sup_{\boldsymbol{\mu}\in\Delta_\Omega} \left\|\boldsymbol{\mu}\boldsymbol{M}^t - \boldsymbol{\pi}\right\|_{\mathsf{TV}} \\
&\overset{(i)}{=} \sup_{\boldsymbol{\mu}\in\Delta_\Omega} \left\|(\boldsymbol{\mu}\boldsymbol{M}^{t-s_{\mathsf{gen}}} - \boldsymbol{\pi})\boldsymbol{M}^{s_{\mathsf{gen}}}\right\|_{\mathsf{TV}} \\
&\overset{(ii)}{\le} \kappa(\boldsymbol{M}^{s_{\mathsf{gen}}}) \sup_{\boldsymbol{\mu}\in\Delta_\Omega} \left\|\boldsymbol{\mu}\boldsymbol{M}^{t-s_{\mathsf{gen}}} - \boldsymbol{\pi}\right\|_{\mathsf{TV}} \\
&\overset{(iii)}{\le} \kappa(\boldsymbol{M}^{s_{\mathsf{gen}}})^{\lfloor t/s_{\mathsf{gen}}\rfloor} \sup_{\boldsymbol{\mu}\in\Delta_\Omega} \left\|\boldsymbol{\mu}\boldsymbol{M}^{t-\lfloor t/s_{\mathsf{gen}}\rfloor s_{\mathsf{gen}}} - \boldsymbol{\pi}\right\|_{\mathsf{TV}} \\
&\overset{(iv)}{\le} \kappa(\boldsymbol{M}^{s_{\mathsf{gen}}})^{(t-s_{\mathsf{gen}})/s_{\mathsf{gen}}}
\end{aligned}$$

where (i) is by definition of $\boldsymbol{\pi}$, (ii) is the contraction property at (7), (iii) is by an inductive argument, and (iv) is by property of the total variation distance. Since $1-\kappa_{\mathsf{gen}} = \frac{1-\kappa(\boldsymbol{M}^{s_{\mathsf{gen}}})}{s_{\mathsf{gen}}} \le \frac{1}{s_{\mathsf{gen}}}$, and from properties of the exponential function, $h(t) \le e\cdot e^{-t(1-\kappa_{\mathsf{gen}})}$, so that for $t > \frac{\ln e/\xi}{1-\kappa_{\mathsf{gen}}}$, $h(t) \le \xi$, hence the upper bound.
For the lower bound, notice that $\forall (i,j)\in\Omega^2$, by sub-additivity of the ℓ_1 norm and by definition of successively $h(t)$ and $t_{\mathsf{mix}}(\xi)$,

$$\left\|\boldsymbol{M}^{t_{\mathsf{mix}}(\xi)}(i,\cdot) - \boldsymbol{M}^{t_{\mathsf{mix}}(\xi)}(j,\cdot)\right\|_{\mathsf{TV}} \le 2h(t_{\mathsf{mix}}(\xi)) \le 2\xi,$$

such that by definition of κ_{gen} and κ,

$$1-\kappa_{\mathsf{gen}} \ge \frac{1-\kappa(\boldsymbol{M}^{t_{\mathsf{mix}}(\xi)})}{t_{\mathsf{mix}}(\xi)} \ge \frac{1-2\xi}{t_{\mathsf{mix}}(\xi)}.$$

The lower bound is also a consequence of (Paulin, 2016, Proposition 3.3, (3.2)).

□

4.2. Proof of Theorem 2

We first report Wolfer and Kontorovich (2019, Lemma D.4), which we will use in our argument.

Lemma 9 (Wolfer and Kontorovich (2019)) *Let $X_1, \dots, X_m \sim (\boldsymbol{M}, \boldsymbol{\mu})$ a d-state ergodic Markov chain. Then, with probability at least $1-\delta$,*

$$\left\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\right\|_\infty \leq 4\mathcal{L}\sqrt{\frac{d}{N_{\min}}},$$

where

$$\mathcal{L} \doteq \arg\min_{t\geq 1}\left\{\left(1+\lceil \ln(2m/t)\rceil_+\right)(d+1)e^{-t} \leq \delta/d\right\} = \mathcal{O}\left(\ln\left(\frac{d\ln m}{\delta}\right)\right),$$

$\widehat{\boldsymbol{M}}$ is the empirical transition matrix of counts defined at (12), and $N_{\min}$ is defined at (11).

In other words, Lemma 9 shows that it is possible to control with high-probability the error in estimating the Markov kernel w.r.t the ℓ_∞ operator norm in terms of *the least number of visits*. Writing for convenience $\kappa_{\mathsf{gen}[S]} \doteq \max_{s\in[S]}\left\{\frac{1-\kappa_s}{s}\right\}$, and for $r \in S$,

$$\mathcal{L}_r \doteq \arg\min_{t\geq 1}\left\{\left(1+\lceil \ln(2m/(tr))\rceil_+\right)(d+1)e^{-t} \leq \frac{\delta}{dS}\right\} = \mathcal{O}\left(\ln\left(\frac{dS\ln m/r}{\delta}\right)\right).$$

Then successively,

$$\begin{aligned}
&\mathbb{P}\left(\left|\widehat{\kappa}_{\mathsf{gen}[S]} - \kappa_{\mathsf{gen}}\right| > \frac{1}{S} + \max_{r\in[S]}\left\{\frac{4}{r}\mathcal{L}_r\sqrt{\frac{d}{N_{\min}^{(r)}}}\right\}\right)\\
&\overset{(i)}{\leq} \mathbb{P}\left(\max_{s\in[S]}\left|\frac{\widehat{\kappa}_s - \kappa_s}{s}\right| > \max_{r\in[S]}\left\{\frac{4}{r}\mathcal{L}_r\sqrt{\frac{d}{N_{\min}^{(r)}}}\right\}\right)\\
&\overset{(ii)}{\leq} \sum_{s=1}^{S}\mathbb{P}\left(\left|\widehat{\kappa}_s - \kappa_s\right| > s\max_{r\in[S]}\left\{\frac{4}{r}\mathcal{L}_r\sqrt{\frac{d}{N_{\min}^{(r)}}}\right\}\right)\\
&\overset{(iii)}{\leq} \sum_{s=1}^{S}\mathbb{P}\left(\left\|\widehat{\boldsymbol{M}}^{(s)} - \boldsymbol{M}^s\right\|_\infty > 4\mathcal{L}_s\sqrt{\frac{d}{N_{\min}^{(s)}}}\right)\\
&\overset{(iv)}{\leq} \sum_{s=1}^{S}\frac{\delta}{S} = \delta,
\end{aligned}$$

where (i) follows from the fact that $\left|\kappa_{\mathsf{gen}} - \kappa_{\mathsf{gen}[S]}\right| \leq 1/S$ and that for $\boldsymbol{\nu}, \boldsymbol{\theta} \in \mathbb{R}^S$ it is the case from sub-additivity of the uniform norm that $\left|\|\boldsymbol{\nu}\|_\infty - \|\boldsymbol{\theta}\|_\infty\right| \leq \|\boldsymbol{\nu} - \boldsymbol{\theta}\|_\infty$; (ii) is an application of the union bound, (iii) stems from the fact that the ℓ_∞ operator norm dominates the distance between Dobrushin contraction coefficients (Fact 5.1), and (iv) is Lemma 9. □

4.3. Proof of Theorem 5

To reach arbitrary precision down to additive error, we explore the first $S = \lceil 2/\varepsilon \rceil$ possible skipping rates, i.e. consider the estimator $\widehat{\kappa}_{\mathsf{gen}\lceil 2/\varepsilon \rceil}$. Then, following the same first steps $(i), (ii), (iii)$ as in the proof of Theorem 2 together with $S > \frac{2}{\varepsilon}$,

$$\mathbb{P}_{\boldsymbol{\pi}}\left(\left|\widehat{\kappa}_{\mathsf{gen}\lceil 2/\varepsilon \rceil} - \kappa_{\mathsf{gen}}\right| > \varepsilon\right) \leq \sum_{s=1}^{\lceil 2/\varepsilon \rceil} \mathbb{P}_{\boldsymbol{\pi}}\left(\left\|\widehat{\boldsymbol{M}}^{(s)} - \boldsymbol{M}^s\right\|_\infty > s\frac{\varepsilon}{2}\right).$$

For each term,

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\pi}}\left(\left\|\widehat{\boldsymbol{M}}^{(s)} - \boldsymbol{M}^s\right\|_\infty > s\frac{\varepsilon}{2}\right) &\overset{(i)}{\leq} \mathbb{P}_{\boldsymbol{\pi}}\left(\left\|\widehat{\boldsymbol{M}}^{(s)} - \boldsymbol{M}^s\right\|_\infty > 4\mathcal{L}_s\sqrt{\frac{d}{N_{\min}^{(s)}}}\right) \\
&\quad + \mathbb{P}_{\boldsymbol{\pi}}\left(N_{\min}^{(s)} < \frac{64\mathcal{L}_s^2 d}{s^2\varepsilon^2}\right) \\
&\overset{(ii)}{\leq} \frac{\delta}{2\lceil 2/\varepsilon \rceil} + \mathbb{P}_{\boldsymbol{\pi}}\left(N_{\min}^{(s)} < \frac{1}{2}\lceil (m-1)/s \rceil \pi_\star\right)
\end{aligned}$$

where (i) stems from the fact that for functions of the sample ϕ and ψ, the chaining argument

$$\mathbb{P}_{\boldsymbol{\pi}}\left(\phi(\boldsymbol{X}) > \varepsilon\right) \leq \mathbb{P}_{\boldsymbol{\pi}}\left(\phi(\boldsymbol{X}) > \psi(\boldsymbol{X})\right) + \mathbb{P}_{\boldsymbol{\pi}}\left(\psi(\boldsymbol{X}) > \varepsilon\right)$$

holds, and (ii) follows from the proof of Theorem 2 at confidence $1 - \delta/2$ for the former summand, and by already setting $m \geq c' \frac{d}{\pi_\star s\varepsilon^2}\widetilde{\ln}\left(\frac{dS}{\delta\pi_\star\varepsilon}\right)$, entailing the sufficient $m \geq \frac{128 d\mathcal{L}_s^2}{\pi_\star s\varepsilon^2}$ for the latter. The remaining error probability, which corresponds to an unreasonable number of visits to the least visited state is controlled for $m \geq c''(\frac{1}{\varepsilon} + \frac{t_{\mathsf{mix}}}{\pi_\star}\ln\frac{d}{\varepsilon\delta})$ as a result of Lemma 10, which in turn is a consequence of Chung et al. (2012, Theorem 3.1). Finally, Paulin (2015, Proposition 3.10) extends the bound to non-stationary chains. □

4.4. Proof of Theorem 6

Previously, in order to estimate κ_{gen} in absolute error, we could stop after computing the first $\lceil 2/\varepsilon \rceil$ ergodic coefficients, but for controlling the approximation error with relative accuracy, the algorithm has to investigate on $S = \lceil c/((1-\kappa_{\mathsf{gen}})\varepsilon) \rceil, c \in \mathbb{R}_+$, which is unknown a priori. The solution is to have $\hat{S}(X)$ depend on $N_{\min}$, such that the algorithm will investigate a larger space as more samples are collected. We define the estimator $\widehat{\kappa}_{\mathsf{gen}[\hat{S}]}$ such that

$$\hat{S} \doteq \lceil \sqrt{N_{\min}/d} \rceil.$$

From the triangle inequality,

$$\begin{aligned}
\left|\widehat{\kappa}_{\mathsf{gen}[\hat{S}]} - \kappa_{\mathsf{gen}}\right| &\leq \left|\widehat{\kappa}_{\mathsf{gen}[\hat{S}]} - \widehat{\kappa}_{\mathsf{gen}\lceil 3/(\varepsilon(1-\kappa_{\mathsf{gen}}))\rceil}\right| \\
&+ \left|\widehat{\kappa}_{\mathsf{gen}\lceil 3/(\varepsilon(1-\kappa_{\mathsf{gen}}))\rceil} - \kappa_{\mathsf{gen}\lceil 3/(\varepsilon(1-\kappa_{\mathsf{gen}}))\rceil}\right| \\
&+ \left|\kappa_{\mathsf{gen}\lceil 3/(\varepsilon(1-\kappa_{\mathsf{gen}}))\rceil} - \kappa_{\mathsf{gen}}\right|.
\end{aligned}$$

It is easy to verify that $\left|\kappa_{\mathsf{gen}\lceil 3/(\varepsilon(1-\kappa_{\mathsf{gen}}))\rceil} - \kappa_{\mathsf{gen}}\right| \leq \frac{(1-\kappa_{\mathsf{gen}})\varepsilon}{3}$. To bound the second term with high-probability, we use Theorem 5 at precision $(1-\kappa_{\mathsf{gen}})\varepsilon/3$ and confidence level $1-\delta/3$. It remains to analyze the first term:

$$\begin{aligned}
\left|\widehat{\kappa}_{\mathsf{gen}\lceil\hat{S}\rceil} - \widehat{\kappa}_{\mathsf{gen}\lceil 3/(\varepsilon(1-\kappa_{\mathsf{gen}}))\rceil}\right| &= \left|\max_{\lceil\hat{S}\rceil}\left\{\frac{1-\widehat{\kappa}_s}{s}\right\} - \max_{\lceil 3/((1-\kappa_{\mathsf{gen}})\varepsilon)\rceil}\left\{\frac{1-\widehat{\kappa}_s}{s}\right\}\right| \\
&\leq \max\left\{\frac{1}{s} : s \in [\hat{S}\ldots\lceil 3/((1-\kappa_{\mathsf{gen}})\varepsilon)\rceil] \cup [\lceil 3/((1-\kappa_{\mathsf{gen}})\varepsilon)\rceil\ldots\hat{S}]\right\} \\
&\leq \max\left\{\frac{1}{\hat{S}}, \frac{(1-\kappa_{\mathsf{gen}})\varepsilon}{3}\right\},
\end{aligned}$$

such that for $m \geq 36\frac{d}{\pi_\star(1-\kappa_{\mathsf{gen}})^2\varepsilon^2}$,

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\pi}}\left(\left|\widehat{\kappa}_{\mathsf{gen}\lceil\hat{S}\rceil} - \widehat{\kappa}_{\mathsf{gen}\lceil 3/(\varepsilon(1-\kappa_{\mathsf{gen}}))\rceil}\right| > \frac{(1-\kappa_{\mathsf{gen}})\varepsilon}{3}\right) &\leq \mathbb{P}_{\boldsymbol{\pi}}\left(\frac{1}{\hat{S}} > \frac{(1-\kappa_{\mathsf{gen}})\varepsilon}{3}\right) \\
&\leq \mathbb{P}_{\boldsymbol{\pi}}\left(\sqrt{\frac{d}{N_{\min}}} > \frac{(1-\kappa_{\mathsf{gen}})\varepsilon}{3}\right) \\
&\leq \mathbb{P}_{\boldsymbol{\pi}}\left(|\pi_\star - \hat{\pi}_\star| > \frac{3}{4}\pi_\star\right),
\end{aligned}$$

where $\hat{\pi}_\star$ is the plug-in estimator for $\pi_\star$ defined in Hsu et al. (2015). Thus for $m \geq c\frac{t_{\mathsf{mix}}}{\pi_\star}\ln\left(\frac{d}{\delta}\right), c \in \mathbb{R}_+$ (Wolfer and Kontorovich, 2019, Theorem 1), this is smaller than $\delta/3$. Finally, the result is extended to non-stationary chains with Paulin (2015, Proposition 3.10). Remark: This expression for $\hat{S}$ confirms the practical choice we proposed at (14).

4.5. Proof of Corollary 7

Combining Theorem 1 and Theorem 6, and choosing $\hat{t}_{\mathsf{mix}} \doteq \frac{1}{1-\widehat{\kappa}^{\times}_{\mathsf{gen}}}$, for the value of m in Theorem 6, with probability at least $1-\delta$,

$$\frac{t_{\mathsf{mix}}}{(1+\varepsilon)(1+\ln 4)} \leq \hat{t}_{\mathsf{mix}} \leq \frac{2t_{\mathsf{mix}}}{(1-\varepsilon)},$$

and setting $\varepsilon = 1/4$ yields the corollary. □

5. Auxiliary facts

The following lemma and facts are proved in the appendix.

Lemma 10 *Let $X_1, \ldots, X_m \sim \boldsymbol{M}$ stationary ergodic Markov chain. For a skipping rate s, and for $m \geq c\ln\frac{d}{\delta\varepsilon}\frac{t_{\mathsf{mix}}}{\pi_\star}, c \in \mathbb{R}_+$,*

$$\mathbb{P}_{\boldsymbol{\pi}}\left(N^{(s)}_{\min} < \frac{1}{2}\lceil(m-1)/s\rceil\pi_\star\right) \leq \frac{\delta}{2d\lceil 2/\varepsilon\rceil}.$$

Fact 5.1 *For two Markov matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$,*

$$|\kappa(\boldsymbol{M}_1) - \kappa(\boldsymbol{M}_2)| \leq \|\boldsymbol{M}_1 - \boldsymbol{M}_2\|_\infty .$$

Fact 5.2 *Let $X_1, \ldots, X_m \sim \boldsymbol{M}$ a stationary ergodic Markov chain with stationary distribution $\boldsymbol{\pi}$ and mixing time at most t_{mix}. Then the mixing time $t_{\mathsf{mix}}^{(s)}$ of the skipped chain for $s \in [m]$,*

$$X_1, X_{1+s}, X_{1+2s}, \ldots, X_{1+\lfloor (m-1)/s \rfloor s} \sim (\boldsymbol{\mu}, \boldsymbol{M}^s),$$

is such that $t_{\mathsf{mix}}^{(s)} \leq \lceil t_{\mathsf{mix}}/s \rceil$.

6. Discussion and future research directions

The present work offers a new perspective on the problem of estimating the mixing properties of a Markov chain, switching the focus from spectral methods and t_{rel} to contraction methods and t_{mix} itself. This offers a first step in determining whether these two statistical problems are of equivalent complexity. The proposed algorithms are primarily of theoretical interest, as they remain computationally intensive both in space and time. Algorithmic optimization of the search over the subset S, for example by leveraging additional properties of κ_{gen} is on our research agenda.

Acknowledgments

This research was supported by the Lynn and William Frankel Center for Computer Science at Ben-Gurion University and by the Israel Science Foundation. The author thanks the anonymous referees for their valuable comments.

References

Pierre Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer-Verlag, 1999.

Russ Bubley and Martin Dyer. Path coupling: A technique for proving rapid mixing in markov chains. In *Proceedings 38th Annual Symposium on Foundations of Computer Science*, pages 223–231. IEEE, 1997.

Kai-Min Chung, Henry Lam, Zhenming Liu, and Michael Mitzenmacher. Chernoff-hoeffding bounds for markov chains: Generalized and simplified. *arXiv preprint arXiv:1201.0559*, 2012.

Richard Combes and Mikael Touati. Computationally efficient estimation of the spectral gap of a markov chain. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3 (1):7, 2019.

RL Dobrushin. Central limit theorem for non-stationary markov chains, i, ii. theory prob. appl. 1, 65-80, 329-383. *English translation*, 1956.

Martin Dyer, Leslie Ann Goldberg, Catherine Greenhill, Mark Jerrum, and Michael Mitzenmacher. An extension of path coupling and its application to the glauber dynamics for graph colorings. *SIAM Journal on Computing*, 30(6):1962–1975, 2001.

Daniel Hsu, Aryeh Kontorovich, David A. Levin, Yuval Peres, Csaba Szepesvri, and Geoffrey Wolfer. Mixing time estimation in reversible markov chains from a single sample path. *Ann. Appl. Probab.*, 29(4):2439–2480, 08 2019. doi: 10.1214/18-AAP1457. URL https://doi.org/10.1214/18-AAP1457.

Daniel J Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible markov chains from a single sample path. In *Advances in neural information processing systems*, pages 1459–1467, 2015.

Daniel Jerison. General mixing time bounds for finite markov chains via the absolute spectral gap. *arXiv preprint arXiv:1310.8021*, 2013.

Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the $l\_\{1\}$ distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, 2018.

Sudeep Kamath and Sergio Verdú. Estimation of entropy rate and rényi entropy rate for markov chains. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 685–689. IEEE, 2016.

David A Levin and Yuval Peres. Estimating the spectral gap of a reversible markov chain from a short trajectory. *arXiv preprint arXiv:1612.05330*, 2016.

David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov chains and mixing times, second edition*. American Mathematical Soc., 2009.

Malwina Luczak. Concentration of measure and mixing for markov chains. *Discrete Mathematics & Theoretical Computer Science*, 2008.

Yann Ollivier. Ricci curvature of markov chains on metric spaces. *Journal of Functional Analysis*, 256:810–864, 2009.

Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20, 2015.

Daniel Paulin. Mixing and concentration by ricci curvature. *Journal of Functional Analysis*, 270 (5):1623–1662, 2016.

Qian Qin, James P Hobert, Kshitij Khare, et al. Estimating the spectral gap of a trace-class markov operator. *Electronic Journal of Statistics*, 13(1):1790–1822, 2019.

Geoffrey Wolfer and Aryeh Kontorovich. Estimating the mixing time of ergodic markov chains. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 3120–3159. PMLR, 2019. URL http://proceedings.mlr.press/v99/wolfer19a.html.

Proof of auxiliary lemmas and facts

Lemma 10 Let $X_1, \dots, X_m \sim \boldsymbol{M}$ stationary ergodic Markov chain. For a skipping rate s, and for $m \geq c \ln \frac{d}{\delta\varepsilon} \frac{t_{\mathsf{mix}}}{\pi_\star}, c \in \mathbb{R}_+$,

$$\mathbb{P}_{\boldsymbol{\pi}}\left(N_{\min}^{(s)} < \frac{1}{2}\lceil (m-1)/s \rceil \pi_\star\right) \leq \frac{\delta}{2d\lceil 2/\varepsilon \rceil}.$$

Proof From the Chernoff-Hoeffding lower tail concentration inequality at Chung et al. (2012, Theorem 3.1), for $X_1, \dots, X_m \sim \boldsymbol{M}$ ergodic over d states, stationary, with mixing time $t_{\mathsf{mix}}(\xi)$ for $\xi \leq 1/8$, and $\eta \in (0,1)$,

$$\mathbb{P}_{\boldsymbol{\pi}}\left(N_i \leq (1-\eta)(m-1)\boldsymbol{\pi}(i)\right) \leq c \exp\left(-\frac{\eta^2 m \boldsymbol{\pi}(i)}{72 t_{\mathsf{mix}}(\xi)}\right), c \in \mathbb{R}_+$$

where we already used the definition of N_i and that by stationarity, $\mathbb{E}_{\boldsymbol{\pi}}\left[\mathbf{1}\left\{X_t = i\right\}\right] = \boldsymbol{\pi}(i)$. The astute reader will notice that our definition of t_{mix} is for $\xi = 1/4$, such that this theorem is not applicable verbatim, however Chung et al. (2012) mentions (at p.3) that it generally holds with $\frac{1-\sqrt{2\xi}}{36 t_{\mathsf{mix}}(\xi)}$ instead of $\frac{1}{72 t_{\mathsf{mix}}(\xi)}$. In our case, we therefore adapt the constant to $c' = \frac{1-\sqrt{1/2}}{36}$, and for $\eta = 1/2$,

$$\mathbb{P}_{\boldsymbol{\pi}}\left(N_i \leq \frac{(m-1)\boldsymbol{\pi}(i)}{2}\right) \leq c \exp\left(-\frac{c'(m-1)\boldsymbol{\pi}(i)}{t_{\mathsf{mix}}}\right). \tag{15}$$

We proceed and apply the above to the different multi-step chains for $s \in [S]$. The mixing time $t_{\mathsf{mix}}^{(s)}$ of each such chain is at most about $\frac{t_{\mathsf{mix}}}{s}$, which is formalized in Fact 5.2.

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\pi}}\left(N_i^{(s)} \leq \frac{1}{2}\lceil (m-1)/s \rceil \boldsymbol{\pi}(i)\right) &\overset{(i)}{\leq} c \exp\left(-\frac{c' \lfloor (m-1)/s \rfloor \boldsymbol{\pi}(i)}{t_{\mathsf{mix}}^{(s)}}\right) \\ &\overset{(ii)}{\leq} c \exp\left(-\frac{c'(m-s-1)\boldsymbol{\pi}(i)}{t_{\mathsf{mix}}}\right)\end{aligned}$$

where (i) is (15) applied to the skipped chain, as it is still the case that $\mathbb{E}_{\boldsymbol{\pi}}\left[\mathbf{1}\left\{X_{1+s(t-1)} = i\right\}\right] = \boldsymbol{\pi}(i)$, and (ii) is Fact 5.2. As a consequence for $m \geq c''(\frac{1}{\varepsilon} + \frac{t_{\mathsf{mix}}}{\boldsymbol{\pi}(i)} \ln \frac{2d\lceil 2/\varepsilon \rceil}{\delta})$, this error probability is smaller than $\frac{\delta}{2d\lceil 2/\varepsilon \rceil}$. Taking a maximum over $i \in \Omega$, yields the lemma. ■

Fact 5.1 For two Markov matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$,

$$|\kappa(\boldsymbol{M}_1) - \kappa(\boldsymbol{M}_2)| \leq \|\boldsymbol{M}_1 - \boldsymbol{M}_2\|_\infty.$$

Proof This is a direct consequence of the sub-additivity of the sup norm.

$$\begin{aligned}2\,|\kappa(\boldsymbol{M}_1) - \kappa(\boldsymbol{M}_2)| &\leq \max_{(i,j)\in\Omega^2} \left| \|\boldsymbol{M}_1(i,\cdot) - \boldsymbol{M}_1(j,\cdot)\|_1 - \|\boldsymbol{M}_2(i,\cdot) - \boldsymbol{M}_2(j,\cdot)\|_1 \right| \\ &\leq \max_{(i,j)\in\Omega^2} \|\boldsymbol{M}_1(i,\cdot) - \boldsymbol{M}_1(j,\cdot) - \boldsymbol{M}_2(i,\cdot) + \boldsymbol{M}_2(j,\cdot)\|_1 \\ &\leq \max_{(i,j)\in\Omega^2} \left(\|\boldsymbol{M}_1(i,\cdot) - \boldsymbol{M}_2(i,\cdot)\|_1 + \|\boldsymbol{M}_1(j,\cdot) - \boldsymbol{M}_2(j,\cdot)\|_1\right) \\ &= 2 \max_{i\in\Omega} \|\boldsymbol{M}_1(i,\cdot) - \boldsymbol{M}_2(i,\cdot)\|_1 \\ &= 2\,\|\boldsymbol{M}_1 - \boldsymbol{M}_2\|_\infty.\end{aligned}$$

■

Fact 5.2 Let $X_1, \dots, X_m \sim \boldsymbol{M}$ a stationary ergodic Markov chain with stationary distribution $\boldsymbol{\pi}$ and mixing time at most t_{mix}. Then the mixing time $t_{\mathsf{mix}}^{(s)}$ of the skipped chain for $s \in [m]$,

$$X_1, X_{1+s}, X_{1+2s}, \dots, X_{1+\lfloor (m-1)/s \rfloor s} \sim (\boldsymbol{\mu}, \boldsymbol{M}^s),$$

is such that $t_{\mathsf{mix}}^{(s)} \leq \lceil t_{\mathsf{mix}}/s \rceil$.

Proof Let t such that $t > \lceil t_{\mathsf{mix}}/s \rceil$, then

$$\left\| \boldsymbol{\mu}(\boldsymbol{M}^s)^t - \boldsymbol{\pi} \right\|_{\mathsf{TV}} \leq \left\| \boldsymbol{\mu}(\boldsymbol{M}^s)^{\lceil t_{\mathsf{mix}}/s \rceil} - \boldsymbol{\pi} \right\|_{\mathsf{TV}} = \left\| \boldsymbol{\mu}\boldsymbol{M}^{t_{\mathsf{mix}}} - \boldsymbol{\pi} \right\|_{\mathsf{TV}} \leq \frac{1}{4},$$

where the first inequality holds as advancing the chain can only move it closer to stationarity (Levin et al., 2009, Exercise 4.2). ■

Algorithm We describe in Algorithm 1 the adaptive version of the procedure that outputs an estimator for the mixing time $\hat{t}_{\mathsf{mix}}$, with the guarantees of Corollary 7, modulo a smoothing parameter λ.

The time complexity of the algorithm of Hsu et al. (2019) for estimating the absolute spectral gap of a reversible chain is of the order of $\mathcal{O}(m + d^3)$. The extension of Wolfer and Kontorovich (2019) that involves the first $S \in \mathbb{N}$ multiplicative reversiblizations of the chain, without considering any form of parallelization, has a computational complexity upper bounded by $\mathcal{O}(S(m + d^3))$. Algorithm 1 has an equivalent time complexity, as computing the Dobrushin contraction coefficient requires $\mathcal{O}(d^3)$.

Interestingly, the complexity of constructing the confidence interval compares favorably with the previous methods. In Hsu et al. (2019) the necessity of computing the *pseudo-inverse* of the empirical transition matrix leads to a complexity of $\mathcal{O}(m + d^3)$. In Wolfer and Kontorovich (2019) computing an interval requires $\tilde{\mathcal{O}}(m + d^2)$ for a reversible chain, and $\mathcal{O}(S(m + d^3))$ in the non-reversible case. In our algorithm, computing the interval can be done in $\mathcal{O}(S(m + d))$.

Function `MixingTime` (d, λ, $(X_1, \ldots, X_m)$) **:**
  return 1 / (1 - `GeneralizedContractionCoeffAdaptive` (d, λ, $(X_1, \ldots, X_m)$))
Function `GeneralizedContractionCoeffAdaptive` (d, λ, $(X_1, \ldots, X_m)$) **:**
  $\mathbf{N} \leftarrow [0]_d$ $N_{\min} \leftarrow m$
  for $t \leftarrow 1$ **to** $m-1$ **do**
    $\mathbf{N}[X_t] \leftarrow \mathbf{N}[X_t] + 1$
  end
  for $i \leftarrow 1$ **to** d **do**
    if $\mathbf{N}[i] < N_{\min}$ **then**
      $N_{\min} \leftarrow \mathbf{N}[i]$
    end
  end
  return `GeneralizedContractionCoeff` (d, λ, $(X_1, \ldots, X_m)$, $\lceil \sqrt{N_{\min}/d} \rceil$)
Function `GeneralizedContractionCoeff` (d, λ, $(X_1, \ldots, X_m)$, S) **:**
  $r_{\mathsf{max}} \leftarrow 0$
  for $s \leftarrow 1$ **to** S **do**
    $\kappa \leftarrow$ `ContractionCoeff`$(d, \lambda, (X_1, X_{1+s}, X_{1+2s}, \ldots, X_{1+\lfloor (m-1)/s \rfloor s}))$
    if $(1-\kappa)/s > r_{\mathsf{max}}$ **then**
      $r_{\mathsf{max}} \leftarrow (1-\kappa)/s$
    end
  end
  return $1 - r_{\mathsf{max}}$
Function `ContractionCoeff` $(d, \lambda, (X_1, \ldots, X_n))$ **:**
  $\mathbf{N} \leftarrow [d\lambda]_d$
  $\mathbf{T} \leftarrow [\lambda]_{d \times d}$
  for $t \leftarrow 1$ **to** $n-1$ **do**
    $\mathbf{N}[X_t] \leftarrow \mathbf{N}[X_t] + 1$
    $\mathbf{T}[X_t, X_{t+1}] \leftarrow \mathbf{T}[X_t, X_{t+1}] + 1$
  end
  $a_{\mathsf{max}} \leftarrow 0$
  for $i \leftarrow 1$ **to** d **do**
    for $j \leftarrow 1$ **to** d **do**
      $a = 0$
      for $k \leftarrow 1$ **to** d **do**
        $a \leftarrow a + |\mathbf{T}[i,k]/\mathbf{N}[i] - \mathbf{T}[j,k]/\mathbf{N}[j]|$
      end
      if $a > a_{\mathsf{max}}$ **then**
        $a_{\mathsf{max}} \leftarrow a$
      end
    end
  end
  return $a_{\mathsf{max}}/2$

Algorithm 1: The estimation procedure outputting $\hat{t}_{\mathsf{mix}}$.